Understanding Social Interpersonal Interaction via Synchronization Templates of Facial Events

Rui Li
College of Computing & Info. Sciences
Rochester Institute of Technology
Rochester, New York 14623
Email: rxlics@rit.edu

Jared Curhan
Sloan School of Management
Massachusetts Institute of Technology
Cambridge, Boston 02139
Email: curhan@mit.edu

Mohammed Ehsan Hoque
Department of Computer Science
University of Rochester
Rochester, New York 14627
Email: mehoque@cs.rochester.edu

Abstract

Automatic facial expression analysis in inter-personal communication is challenging. Not only because conversation partners' facial expressions mutually influence each other, but also because no correct interpretation of facial expressions is possible without taking social context into account. In this paper, we propose a probabilistic framework to model interactional synchronization between conversation partners based on their facial expressions. Interactional synchronization manifests temporal dynamics of conversation partners' mutual influence. In particular, the model allows us to discover a set of common and unique facial synchronization templates directly from natural interpersonal interaction without recourse to any predefined labeling schemes. The facial synchronization templates represent periodical facial event coordinations shared by multiple conversation pairs in a specific social context. We test our model on two different dyadic conversations of negotiation and job-interview. Based on the discovered facial event coordination, we are able to predict their conversation outcomes with higher accuracy than HMMs and GMMs.

Introduction

Facial expression plays a key role in human communication and relationships, as it is our direct and naturally preeminent means of communicating and understanding affective state and intentions. There has been significant progress in automatic algorithms for facial expression recognition on both static images and video clips (Pantic and Rothkrantz 2000; Zhao and Pietikainen 2007; Zhong et al. 2012; Liu et al. 2014). On the other hand, automatically interpreting facial expressions in interpersonal interaction is still challenging. One difficulty is to capture the temporal dynamical mutual influence between conversation partners. Social psychology studies show that there is spontaneous mutual influence in social interactions, usually without conscious awareness (Shockley, Santana, and Fowler 2003; Singer et al. 2006). Namely, as active participants, conversation partners constantly adjust their facial expression in response to feedback from each other. Another challenge is that it heavily depends on the social context to correctly decode a person's facial expression. The same facial expression can convey different affective information in different contexts of social interaction (Singer et al. 2006; Tamietto and de Gelder 2010; Barrett, Mesquita, and Gendron 2011). These factors motivate us to investigate facial expressions from the perspective of interactional synchrony.

Interactional synchrony refers to periodic temporal coordination between people's nonverbal behaviors in social interaction (Schmidt et al. 2012). It describes the intrinsic mutual influence of two-way communication in flux. In particular, facial expression synchrony has been found strongly indicative for social signals including (dis-)agreement, empathy, hostility, and any other attitude towards others that cannot be expressed using just words (Gratier 2004; Curhan and Pentland 2007; Dunbar et al. 2014). Besides, face-to-face (FtF) interaction, we also focus on interpersonal interaction via video-conferencing (VC) platforms. The explosive growth of the Internet and VC signifies the importance of understanding the social behaviors in this medium. Moreover, VC communication efficiency is hindered by some drawbacks including the limited view of the person, disengaged eye contact, and occasional interruptions resulting from network latency. Online analysis and interpretation of facial expression synchronization can help to mitigate these deficiencies.

In this study, we investigate pairs of conversation partners' facial expression synchronization while they are engaging in VC and FtF communication. In order to represent interactional synchrony, we develop a novel probabilistic framework to discover and summarize common and unique facial synchronization templates shared among multiple conversation pairs, as shown in Figure 1. The facial synchronization templates are described by synchronized facial events periodically displayed by the conversation partners.

A facial synchronization template essentially characterizes two levels of statistical regularities of conversation pairs' facial expressions.

- At the individual level, a template consists of a similar facial event displayed by the individuals who play the same social role (e.g., recruiters) in the conversations.
- At the dyadic level, a template describes a particular temporal coordination of conversation partners' facial events that is periodically exhibited in the interaction.

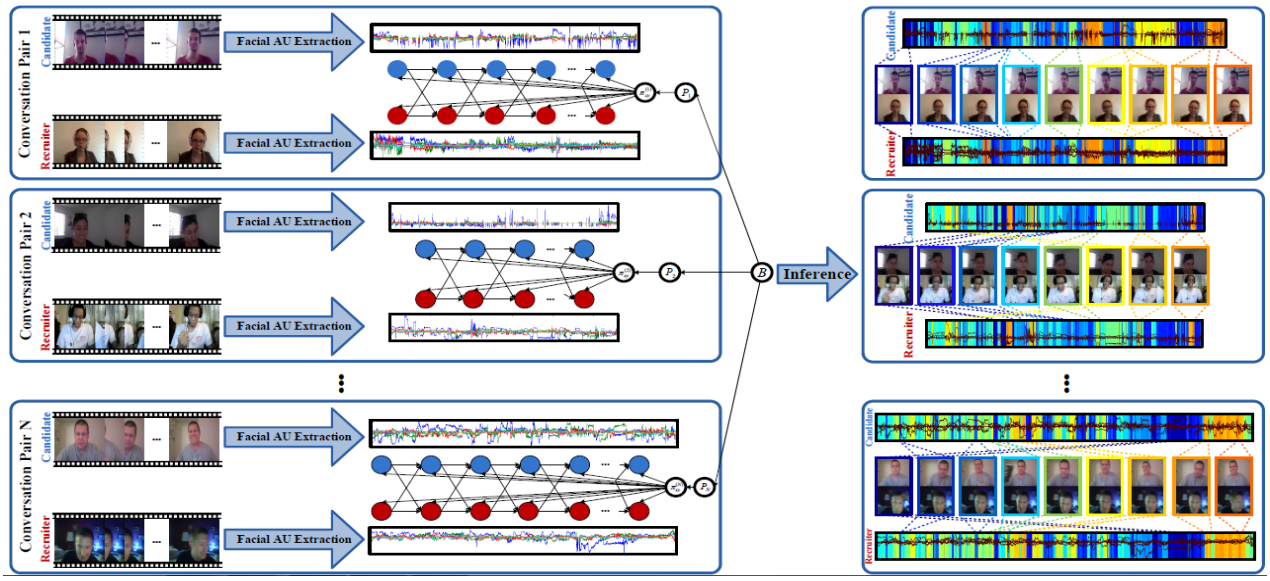

Figure 1: Overview of our approach. From left to right, facial action units (AUs) are extracted from the videos of conversation pairs, and then transformed into time series as the input of our model. The time series are the first 6 principal components from 20 facial AUs evolving over time. A conversation pair's time series are described by a coupled hidden Markov model to capture the interdependence while allowing each time series to maintain its own dynamic process. A beta process prior is used to summarize the facial synchronization templates shared across multiple conversation pairs (details are in Figure 2). The right panel demonstrates the inference output from the model. A conversation pair's facial time series are decomposed into a number of segments (color coded). These segments correspond to a set of globally shared facial synchronization templates. The color coding superposed on the time series indicates the occurrence of the shared templates (e.g., dark blue segments of the time series highlight the same template shared across the conversation pairs). The video frames as the instantiations of each synchronization template are also illustrated in the same color coding scheme. Note that the synchronization characterizes periodical temporal coordination of the conversation partners' facial events. Our model allows each conversation pair to display a different subset of the globally shared synchronization templates, and to switch between them in a unique manner.

We apply our approach to study the interpersonal interaction in two social contexts: VC-based negotiation and FtF interview. Based on the discovered social-role specific facial events and their synchronization templates' occurrence frequencies, we are able to predict the conversation outcomes.

Related Work

Social signals bear critical information of decision making, as they include (dis-)agreement, empathy, hostility, and any other attitude towards others that cannot be expressed using just words. Socially aware computation aims at decoding social signals by automatically interpreting various types of non-linguistic behaviors (Pentland 2005).

Facial expression recognition is approached largely in two steps. The first step is facial representation. Studies at this level try to derive a set of features from original facial images or videos to effectively represent a variety of facial changes (Zhao and Pietikainen 2007; Liu et al. 2014). The second step is to correctly categorize facial expressions based on the representations (Pantic and Rothkrantz 2000). In particular, Facial Action Coding System (FACS) is widely used to study spontaneous facial expression, because of its descriptive power. FACS parses the visible effects of facial muscle activation into action units (AUs). Each AU corresponds to one ore more facial muscle movements. The considerable progress in automatic algorithms to recognize FACS AUs provides solid foundation for facial expression recognition (Tong, Liao, and Ji 2007; Littlewort et al. 2011; Li, Curhan, and Hoque 2015).

Besides recognizing the set of facial expressions corresponding to universal emotion classes, social psychology studies emphasize that accurately interpreting facial expressions depends upon the social context in which they are displayed. A wide variety of information has been exploited to predict social roles. Social roles (e.g., attacker, defender) are recognized based on identifying group activities in multi-person scenes (e.g., sports) (Lan, Sigal, and Mori 2012). A reference role is specified in an event, and other roles are recognized based on the appearance and tempo-spatial features relative to this reference role (Ramanathan, Yao, and Fei-Fei 2013). Progression has also been made on social relation recognition and interaction understanding (Ding and Yilmaz 2011). This work recognizes social relationships based on appearance features, visual concept attributes and scene context.

Interactional synchrony characterizes temporal dynamics of context-specific mutual influence manifested in natural social interactions. Studies suggest that speakers' and lis-

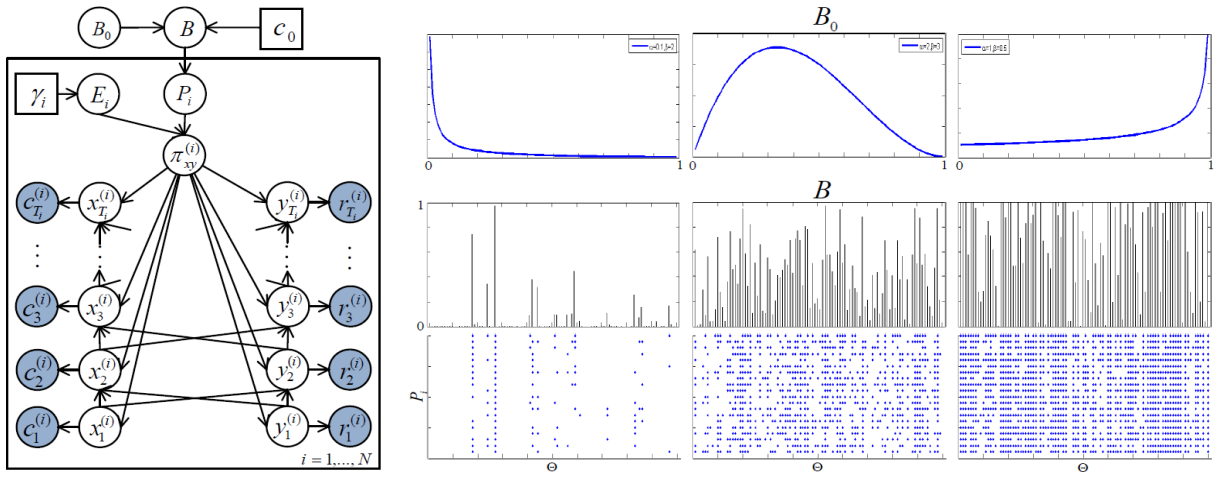

Figure 2: Left: the probabilistic graphical model. Right: three different parameterizations of Beta-Bernoulli processes. The first row are B_0's projections on the probability dimension $[0, 1]$. The second row illustrates B as discrete realizations of a beta process $BP(c_0, B_0)$. They are discrete random measure on the space of synchronization templates Θ. The stick height b_k represents the probability of template θ_k to be displayed by conversation pairs, and its location represent the template in the space. The third row illustrates P_i as a Bernoulli measure given beta process B. In each case, a P_i is a binary row vector of Bernoulli random variables $p_k^{(i)}$s with blue dots denoting 1s and blanks denoting 0s at $\theta_{1:K}$. An element $p_k^{(i)}$ denotes whether a particular synchronization template θ_k is displayed by conversation pair i. So θ_ks with higher sticks tend to generate more Bernoulli realizations. This represents that popular templates tend to be shared by more conversation pairs.

teners' nonverbal behaviors contain rhythms that are not only correlated in time but also exhibited phase synchronization (Bernieri 1988; Schmidt et al. 2012). In a synchronic interaction, nonverbal behaviors of the individuals are coordinated to the rhythms and forms of verbal expressions. Such subtle interpersonal coordination occurs spontaneously without conscious awareness. In particular, this notion helps researchers to decode social signals in a specific social context, and understand inter-personal and group dynamics (Gratier 2004; Shockley, Santana, and Fowler 2003; Dunbar et al. 2014).

The probabilistic Framework

We model time-evolving mutual influence of facial expressions between conversation partners via interactional synchrony. Since conversation partners constantly adjust their facial expression in response to feedback from each other, we describe a pair of conversation partners' facial expressions as two inter-dependent stochastic processes by coupling hidden state variables of two hidden Markov models (HMMs). At each time step an individual's facial expression depends on both his/her own previous one and the partner's previous facial expression, as in Figure 1.

Given a social context (e.g., negotiation), the facial expressions of multiple conversation pairs are distinct yet related. We use Beta process prior to summarize the salient facial synchronization templates shared among multiple pairs. This prior enables us to profile a set of stereotypical and idiosyncratic synchronization templates shared among multiple pairs of expression sequences, as in Figure 2.

Dynamic Likelihoods

Since a pair of conversation partners' facial expressions are assumed to be inter-dependent and meanwhile maintaining their own internal dynamic, two HMMs are coupled via a matrix of conditional probabilities between their hidden state variables, as depicted in Figure 2.

We denote the observations of the i^{th} conversation pair's facial expression sequences as $O_i = \{c_{1:T_i}^{(i)}, r_{1:T_i}^{(i)}\}$, where $c_{1:T_i}^{(i)}$ are the facial expressions displayed by one party, and $r_{1:T_i}^{(i)}$ are the other's (e.g., candidate Vs. recruiter). We further define $S_i = \{x_{1:T_i}^{(i)}, y_{1:T_i}^{(i)}\}$ as the ith pair's hidden state sequences, and $x_{1:T_i}^{(i)}$ and $y_{1:T_i}^{(i)}$ are the respective hidden state sequences of the conversation pair. The state transition probabilities are thus denoted as

$$x_{t+1}^{(i)} | x_t^{(i)}, y_t^{(i)} \sim Mult(\pi_{x_t^{(i)}, y_t^{(i)}}^{(i)}) \quad (1)$$

$$y_{t+1}^{(i)} | y_t^{(i)}, x_t^{(i)} \sim Mult(\pi_{x_t^{(i)}, y_t^{(i)}}^{(i)}) \quad (2)$$

The emission distributions are defined as two respective sets of normal distributions indexed by the corresponding hidden states:

$$c_t^{(i)} | x_t^{(i)} \sim Norm(\mu_{x_t^{(i)}}, \Sigma_{x_t^{(i)}}) \quad (3)$$

$$r_t^{(i)} | y_t^{(i)} \sim Norm(\mu_{y_t^{(i)}}, \Sigma_{y_t^{(i)}}) \quad (4)$$

Therefore, a shared facial synchronization template θ_k can be further denoted as a unique combination of two normal distributions: $\theta_k = ((\mu_k^{(x)}, \Sigma_k^{(x)}), (\mu_k^{(y)}, \Sigma_k^{(y)}))$. Specifically,

Figure 3: Nine facial synchronization templates are illustrated via eight pairs of conversation partners. The social context is a candidate-recruiter negotiation on compensation package via a video-conferencing platform. The matrix consists of instantiations (example frames) of the facial synchronization templates learned by our model. In this matrix, each column contains the instantiations of one particular synchronization template (color coded) shared among the conversation pairs, and each row is one particular conversation pair. In each synchronization template, the candidates display a similar facial event, and the recruiters also display a similar facial event. Such particular facial coordination between the partners is periodically manifested during the conversation. The templates demonstrated here are the most popular ones displayed among the conversation pairs.

$\{(\mu_k^{(x)}, \Sigma_k^{(x)})\}_k$ and $\{(\mu_k^{(y)}, \Sigma_k^{(y)})\}_k$ are the unique facial expression events indexed by $\{x_{T_i}^{(i)}\}_i$ and $\{y_{T_i}^{(i)}\}_i$.

Prior for Synchronization Templates

We propose to use Beta-Bernoulli process prior to relate facial expressions exhibited by conversation pairs (Li et al. 2016). This prior allows us to learn the number of globally shared synchronization templates from multiple pairs of facial expression sequences. Furthermore, each pair can exhibit only a subset of the shared facial synchronization templates, and switch between them in a unique manner.

As shown in Figure 2, let B_0 denote a fixed continuous base measure on a space $\Theta \times [0, 1]$. Θ represents a space of all the possible synchronization templates of facial expressions displayed in a given social context. For multiple conversation pairs to share these templates, let B denote a discrete realization of a Beta process given the prior $BP(c_0, B_0)$. It represents a discrete random measure on the facial synchronization templates $\{\theta_k\}$ displayed among the multiple conversation pairs. Let P_i denote a Bernoulli measure given the beta process B. P_i is a binary vector of Bernoulli random variables representing whether a particular synchronization template θ_k displayed in the facial expression sequences of conversation pair i. This construction can be formulated as follows:

$$B|B_0 \sim BP(c_0, B_0) \quad B = \sum_k b_k \delta_{\theta_k} \tag{5}$$

$$P_i|B \sim BeP(B) \quad P_i = \sum_k p_k^{(i)} \delta_{\theta_k} \tag{6}$$

The above equations show that B is associated with a set of countable number of synchronization templates $\{\theta_k\}$ drawn from Θ as well as their corresponding probability masses $\{b_k\}$. The combination of these two variables characterizes how likely a particular template is shared by the conversation pairs. P_i is a Bernoulli process realization from the random measure B where $p_k^{(i)}$ is a binary random variable denoting whether conversation pair i exhibits template θ_k, given the probability mass b_k.

Based on the above formulations, for $k = 1, ..., K$ we readily define $\{(\theta_k, b_k)\}$ as a set of facial synchronization templates shared among the conversation pairs and $\{(\theta_k, p_k^{(i)})|p_k^{(i)} = 1\}$ as a subset of templates discovered in pair i's facial expression sequences.

The transition distribution $\pi_{xy}^{(i)}$ of the coupled HMMs governs the transitions between the i^{th} pair's subset of templates $\{(\theta_k, p_k^{(i)})|p_k^{(i)} = 1\}$. It is determined by the element-wise multiplication between the subset $\{p_k^{(i)}\}$ and the gamma-distributed random variables $\{e_k^{(i)}\}$:

$$e_k^{(i)}|\gamma_i \sim \Gamma(\gamma_i, 1) \qquad \pi_{xy}^{(i)} \propto E_i \bigotimes P_i \tag{7}$$

where $E_i = [e_1^{(i)}, ...e_K^{(i)}]$. So the effective dimensionality of $\pi_{xy}^{(i)}$ is determined by P_i.

Inference

A Markov chain Monte Carlo method is applied to do posterior inference and parameter learning given the prior Beta

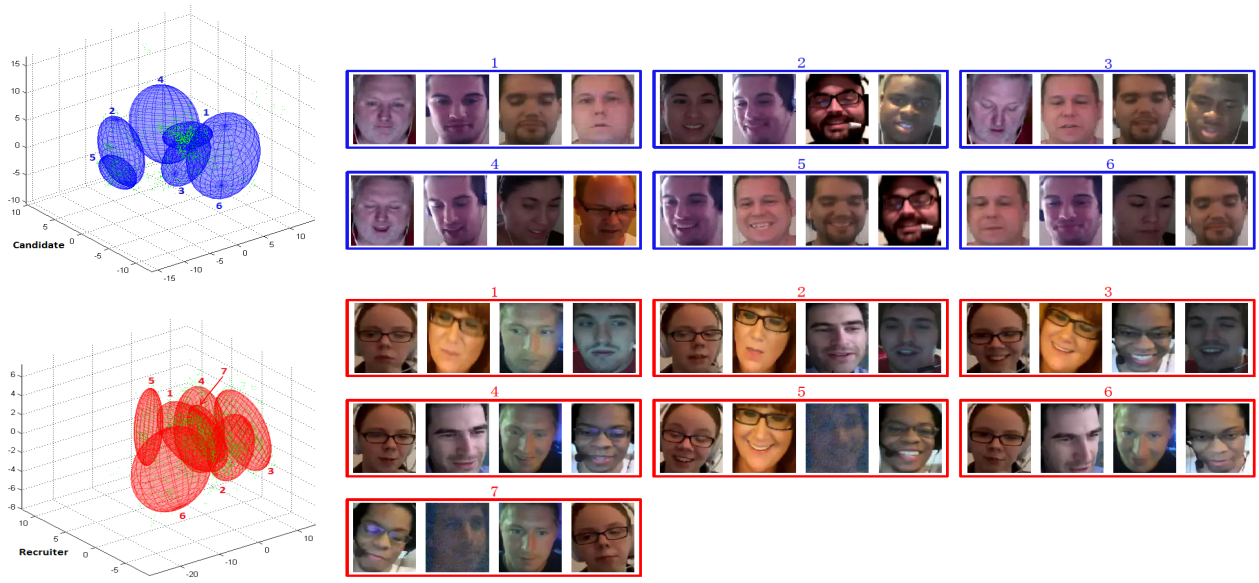

Figure 4: Visualization of the facial event distributions estimated from the conversation partners. Left: the distributions are visualized in the space of the first 3 PCs, and each data point (green) represent a facial expression in one video frame. 6 and 7 Gaussian clusters are learned from the candidates (blue) and the recruiters (red), respectively. Right: we demo four facial event exemplars sampled for each corresponding numbered Gaussian cluster. although some facial events (neutral face, speaking, mild/big smile,...) are similar between candidates and recruiters, there are some subtle difference. For example, the recruiters' sixth facial event exhibiting subtle smiles while looking into the screens is not displayed by the candidates. A facial synchronization template can thus be described as a particular combination of two respective facial events from the two groups.

process (Li, Shi, and Haake 2013). We develop a Gibbs sampling solution to iteratively sample the marginalized Beta-Bernoulli processes of the model given the hidden states and the observations, and then use dynamic programming to update the hidden states given the marginalized prior and the observations.

The following posterior distribution is used to infer whether template $\{(\theta_k, p_k^{(i)})\}$ is displayed in conversation pair i where $i = 1, \ldots, N$, if θ_k is already instantiated by conversation pairs:

$$\begin{aligned} &p(p_k^{(i)} = 1 | P_{-ik}^{(1:N)}, O_i, \theta_{-k}, E_i, c_0) \\ &\propto p(O_i | P_i, E_i, \theta_{-k}) p(p_k^{(i)} = 1 | P_{-ik}^{(1:N)}, c_0) \end{aligned} \tag{8}$$

where $P_{-ik}^{(1:N)}$ represent the binary matrix of $[P_1, P_2, .., P_N]$ except binary vector P_i's k^{th} template. $p(p_k^{(i)} | P_i^{(-k)}, c_0)$ is sampled using Indian buffet process which is a marginalized construction of a Beta-Bernoulli process:

$$\begin{aligned} &p(p_k^{(i)} = 1 | P_{-ik}^{(1:N)}, c_0) \\ &= \int_0^1 p(p_k^{(i)} | b_k) p(b_k | P_{-ik}^{(1:N)}) db_k = \frac{n_{-ik} + \frac{c_0}{K}}{N + \frac{c_0}{K}} \end{aligned} \tag{9}$$

A $Poisson(\frac{c_0}{N})$ distributed number of new templates associated with each pair are drawn in every iteration, and the templates that have no realizations will be deleted.

P_i and E_i determine $\pi_{xy}^{(i)}$ according to Equation 7. Given transition distributions $\pi_{xy}^{(i)}$, shared templates $\{\theta_k\}$, and observed facial expression sequences $c_{1:T_i}^{(i)}$ and $r_{1:T_i}^{(i)}$, within a message passing algorithm, we compute the backward messages:

$$m_{t+1,t}(S_t^{(i)}) \propto p(O_{t+1:T_i}^{(i)} | S_t^{(i)}, \pi_{xy}^{(i)}, \{\theta_k\}) \tag{10}$$

to update the hidden state sequences $S_{1:T_i}^{(i)}$ by sampling from:

$$\begin{aligned} &p(S_t^{(i)} | S_{t-1}^{(i)}, O_{1:T_i}^{(i)}, \pi_{xy}^i, \{\theta_k\}) \\ &\propto \pi_{S_{t-1}^{(i)}}^{(i)}(S_t^{(i)}) N(O_t^{(i)}; \mu_{S_t^{(i)}}^{(i)}, \Sigma_{S_t^{(i)}}^{(i)}) m_{t+1,t}(S_t^{(i)}) \end{aligned} \tag{11}$$

Since $\{\theta_k\}$ parameterize a set of normal distributions, we couple them with a normal inverse-Wishart distribution as a conjugate prior. Both $(\mu_k^{(x)}, \Sigma_k^{(x)})$ and $(\mu_k^{(y)}, \Sigma_k^{(y)})$ are sampled from:

$$\begin{aligned} \Sigma_k &\sim IW(\sum_{i=1}^{N} n_i^{(i)} + n_0, \Sigma_{O_i}^{(k)} + \Sigma_0) \\ \mu_k | \Sigma_k &\sim N(\mu_{O_i}^{(k)}, \Sigma_k) \end{aligned} \tag{12}$$

Experiments and Results

We extract 20 facial AUs from the conversation videos using the Computer Expression Recognition Toolbox (CERT) whose performance is extensively tested (Littlewort et al. 2011). After transforming the AU intensities with the principal component analysis (PCA), we adopt the first 6 PCs, which account for about 97% of the data variance, to represent the facial expression sequences O, as in Figure 1.

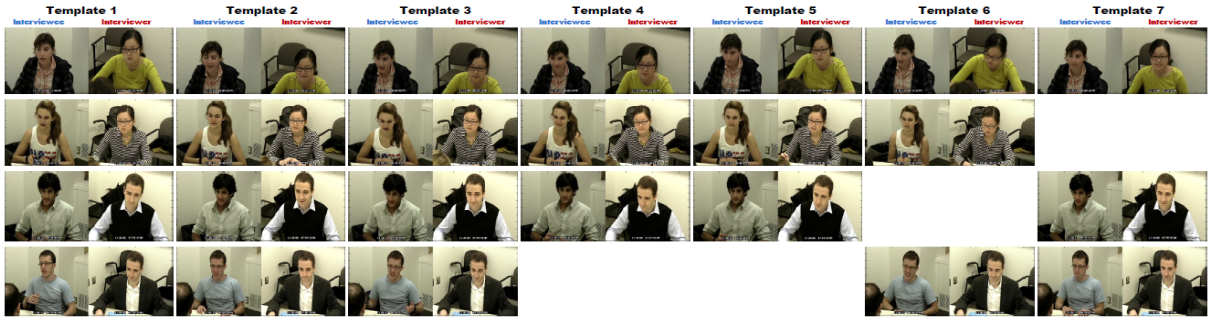

Figure 5: Seven facial synchronization templates are illustrated via four pairs of conversation partners. The social context is a job interview of face-to-face interaction. The matrix consists of example frames of the facial synchronization templates learned by our model from the conversation videos.

We compare our model with two generative models: canonical hidden Markov models (HMMs) and Gaussian mixture models (GMMs). We learn a HMM and a GMM for each social role group using the expectation-maximization algorithm. Their cardinality numbers are determined via five-fold cross-validation.

We apply quadratic discriminant analysis (QDA) for the conversation outcome prediction to test the performance. In particular, we use cross-validation scheme to recursively assign random 60% conversation videos of the two datasets into the training set and the rest into the testing set for the performance comparison. We represent a testing conversation pair's negotiation process O^* by the occurrence frequencies of the synchronization templates $(N_1/N^*, \ldots, N_K/N^*)^T$, where $N_k = \sum_{i=1}^{N^*} \delta(S^* = \theta_k)$. The predictive synchronization templates S^* are computed from $p(S^*|O^*, \{\theta_k\}, \pi_{xy})$. $\{\theta_k\}$ and π_{xy} are inferred in the training phase as described above. We specify a noninformative uniform base measure B_0 as in Figure 2, and compute the posterior by initializing 4 chains of 10,000 sampling iterations on the training data. We then perform the Gelman-Rubin diagnostic (Brooks and Gelman 1998) to assess convergence by calculating the within-chain and between-chain variance on the MCMC samples of the posterior.

Datasets

We validate our method using two datasets collected from two different social contexts.

The negotiation 242 Mechanical Turkers participate in the study. Participants are informed that their negotiations would be recorded and that the study's purpose is to investigate negotiation skills. The data collected from 150 of the Turkers is available for further analysis. The remaining Turkers either had damaged videos or lacked post-questionnaire data. The negotiators interact with each other through a computer-mediated platform based on a browser-based VC system. The system can capture and analyze the video stream in the cloud. We implement the functionality to transfer audio and video data every 30 seconds to prevent data loss and dynamically adapt to variant network latency.

A recruitment case involves a scenario in which a candidate who already has an offer needs to negotiate the compensation package with the recruiter. The candidates and the recruiters need to reach an agreement on eight issues related to salary, job assignment, location, vacation time, bonus, moving expense reimbursement, starting date, and health insurance (Curhan and Pentland 2007). Each negotiation issue offers 5 possible options for resolution. Each option is associated with a specific number of points for each party. The goal of the negotiators is to maximize the total points they can possibly earn (e.g., the 5 optional offers on salary issue range from 65K to 45K. Candidate receives maximum points if he/she could settle with salary of 65k whereas recruiter loses maximum points, whereas recruiter receives maximum points with 45K.).

Participants are randomly formed into 75 pairs, with one member of each pair randomly assigned the role of candidate and the other assigned the role of recruiter. Participants coordinate with their partners to choose the locations and times for the VC-based negotiation, so they may interact in convenient and comfortable circumstances. After both participants provide consent, a button appears that leads each individual to the correct video chat room, which signals that the two can speak with each other. The participants then proceed to play out the scenario outlined in their instructions. Recording begins the moment the two participants connect and are able to see each other and stops when one participant hangs up upon completion of the negotiation. Participants are free to offer whatever information, arguments, and proposals they wish, although they may not exchange their confidential instructions.

The interviews are conducted by two professional career counselors in a room equipped with two wall-mounted cameras. The cameras capture the facial expressions during the interview. The 90 student participants are randomly assigned to the counselors. During each interview session, the counselor asks interviewees five questions (Naim et al. 2016).

VC-based Negotiation

Exploratory interpretations of these templates help us to evaluate their significance. In Figure 3, Template 1 represents a case in which the partners exhibit neutral facial expressions to each other. Template 2 demonstrates that re-

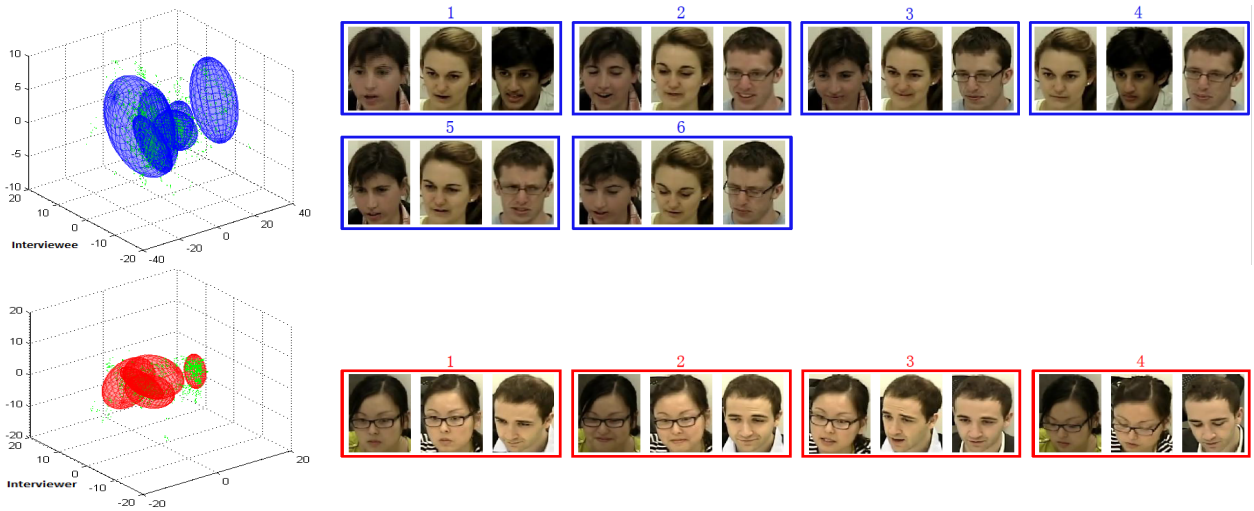

Figure 6: Visualization of the facial event distributions estimated from the job interview scenario.

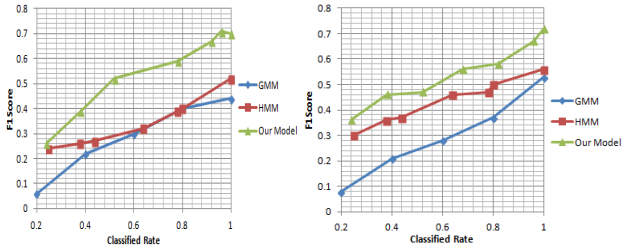

Figure 7: F1 score curves for VC-based negotiation (left) and FtF interview (right) outcome predictions by the QDAs based on the template occurrence frequencies from the three models.

cruiters hold the conversation turn, and the candidates display neutral faces. In Template 3, recruiters guide the conversation, and candidates respond with mild smiles.

The synchronization templates manifest the periodical temporal coordination of the conversation partners' facial events. Figure 4 shows the Gaussian emission distributions of the role-specific facial events as described in Equation 3 and 4. Since the data clusters are visualized in the first 3 principal component space, some separations may not be obvious. The facial event exemplars demonstrate some subtle difference between the candidates and the recruiters. The visualized facial event exemplars in Figure 4 demonstrate some subtle difference between the two social role groups. The recruiters' facial event 1 appears distracted (or thinking) rather than just neutral. This facial event is not displayed by the candidates. Facial event 6 shows that the recruiters display mild smiles and look into the screen, whereas the candidates tend to look down with mild smiles.

We use t-test to measure correlations between the occurrence frequencies of the discovered facial synchronization templates and the points earned by candidates and recruiters, respectively. In Table 1, we show three significant correlations between the templates and candidates'/recruiters'

Synch. Template	Facial Event	Corr. ($p < .05$)
Template 3	Can. Event 2	$\beta = .237$ (r)
	Rec. Event 2	$t(73) = 1.34$
Template 8	Can. Event 3	$\beta = -.182$ (c)
	Rec. Event 5	$t(73) = 1.57$
Template 9	Can. Event 5	$\beta = .209$ (r)
	Rec. Event 1	$t(73) = 2.21$

Table 1: Significant correlations between occurrence frequencies of the facial synchronization templates and the points earned by either social role group. The second column are the facial events of the corresponding templates. The positive correlation of Template 3 and Template 9 with recruiters' points (labeled by r) suggests that recruiters earned more points in interactions when the two templates are displayed more frequently. The negative correlations of Template 8 with candidates' points (labeled by c) suggests that candidates earned fewer points when this template occurs more frequently.

points. The three templates are demonstrated in Figure 3. The facial events in Template 3 are candidates' mild smile and recruiters' speaking, and the facial events in Template 9 are candidates' big smile and recruiters' neutral face. The positive correlations between these two templates and recruiters' points suggests that recruiters earn more points if the scenario occurs frequently that candidates are smiling when they are not. The facial events in Template 8 are recruiters' smile and candidates are speaking. The negative correlations between the template and candidates' points suggests that candidates earn fewer points if the scenario occurs frequently that recruiters are smiling while they are not.

Conclusions

We present a novel probabilistic framework to automatically discover a set of temporal facial synchronization templates.

These templates are shared among conversation pairs in the same social context. These templates profile the periodic temporal coordination of the facial events. Based on the occurrence frequencies of the synchronization templates, we are able to predict the conversation outcomes.

Acknowledgments

This work was supported in part by Grant W911NF-15-1-0542 with the US Defense Advanced Research Projects Agency (DARPA) and the Army Research Office (ARO).

References

Barrett, L. F.; Mesquita, B.; and Gendron, M. 2011. Context in emotion perception. *Journal of Current Directions in Psychological Science* 20(5):286–290.

Bernieri, F. J. 1988. Coordinated movement and rapport in teacher-student interactions. *Journal of Nonverbal Behavior* 12(2):120–138.

Brooks, S. P., and Gelman, A. E. 1998. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics* 7(4):434–455.

Curhan, J., and Pentland, A. 2007. Thin slices of negotiation: Predicting outcomes from conversational dynamics within the first 5 minutes. *Journal of Applied Psychology* 92:802–811.

Ding, L., and Yilmaz, A. 2011. Inferring social relations from visual concepts. In *Proceedings of the Thirteenth International Conference on Computer Vision*, 699–706. IEEE Computer Society.

Dunbar, N. E.; Jensen, M. L.; Tower, D. C.; and Burgoon, J. K. 2014. Synchornization of nonverbal behaviors in detecting mediated and non-mediated deception. *Journal of Nonverbal Behavior* 38:355–376.

Gratier, M. 2004. Expressive timing and interactional synchrony between mothers and infants: Cultural similarites, cultural differences, and the immigration experience. *Journal of Cognition Development* 18:533–554.

Lan, T.; Sigal, L.; and Mori, G. 2012. Social roles in hierarchical models for human activity recognition. In *Proceedings of the Twenty-fifth Conference on Computer Vision and Pattern Recognition*, 4321–4328. IEEE Computer Society.

Li, R.; Shi, P.; Pelz, J.; Alm, C. O.; and Haake, A. R. 2016. Modeling eye movement patterns to characterize perceptual skill in image-based diagnostic reasoning processes. *Journal of Computer Vision and Image Understanding* 151(4):138–152.

Li, R.; Curhan, J.; and Hoque, M. E. 2015. Predicting video-conferencing conversation outcomes based on modeling facial expression synchronization. In *Proceedings of the Eleventh International Conference on Automatic Face and Gesture Recognition*, 1–6. IEEE Computer Society.

Li, R.; Shi, P.; and Haake, A. 2013. Image understanding from experts' eyes by modeling perceptual skills of diagnostic reasoning processes. In *Proceedings of the Twenty-sixth Conference on Computer Vision and Pattern Recognition*, 2187–2194. IEEE Computer Society.

Littlewort, G.; Whitehill, J.; Wu, T.; Fasel, I.; Frank, M.; Movellan, J.; and Bartlett, M. 2011. The computer expression recognition toolbox (cert). In *Proceedings of the Tenth International Conference on Automatic Face and Gesture Recognition*, 1–6. IEEE Computer Society.

Liu, M.; Shan, S.; Wang, R.; and Chen, X. 2014. Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In *Proceedings of the Twenty-seventh Conference on Computer Vision and Pattern Recognition*, 1749–1756. IEEE Computer Society.

Naim, I.; Tanveer, M. I.; Gildea, D.; and Hoque, M. E. 2016. Automated analysis and prediction of job interview performance. *IEEE Transactions on Affective Computing* 1(99):434–455.

Pantic, M., and Rothkrantz, L. J. 2000. Automatic analysis of facial expressions: The state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22(12):1424–1445.

Pentland, A. 2005. Socially aware computation and communication. *Journal of Computer* 38(3):33–40.

Ramanathan, V.; Yao, B.; and Fei-Fei, L. 2013. Social role discovery in human events. In *Proceedings of the Twenty-sixth Conference on Computer Vision and Pattern Recognition*, 2475–2482. IEEE Computer Society.

Schmidt, R. C.; Morr, S.; Fitzpatrick, P.; and Richardson, M. J. 2012. Measuring the dynamics of interactional synchrony. *Journal of Nonverbal Behavior* 36(1):263–279.

Shockley, K.; Santana, M. V.; and Fowler, C. A. 2003. Mutual interpersonal postural constraints are involved in cooperative conversation. *J. Exp. Psychol. Hum. Percept. Perform.* 29(2):326–332.

Singer, T.; Seymour, B.; O'Doherty, J. P.; Stephan, K. E.; Dolan, R. J.; and Frith, C. D. 2006. Empathic neural responses are modulated by the perceived fairness of others. *Nature* 439(1):466–469.

Tamietto, M., and de Gelder, B. 2010. Neural bases of the non-conscious perception of emotional signals. *Journal of Nature Reviews Neuroscience* 11(1):697–709.

Tong, Y.; Liao, W.; and Ji, Q. 2007. Facial action unit recognition by exploting their dynamic and semantic relationships. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29(10):1683–1699.

Zhao, G., and Pietikainen, M. 2007. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29(6):915–928.

Zhong, L.; Liu, Q.; Yang, P.; Liu, B.; Huang, J.; and Metaxas, D. N. 2012. Learning active facial patches for expression analysis. In *Proceedings of the Twenty-fifth Conference on Computer Vision and Pattern Recognition*, 2562–2569.